

## Kennesaw State University DigitalCommons@Kennesaw State University

---

### Faculty Publications

---

9-2012

# An Item Response Curves Analysis of the Force Concept Inventory

Gary A. Morris  
*Valparaiso University*

Nathan Harshman  
*American University*

Lee Branum-Martin  
*University of Houston*

Eric Mazur  
*Harvard University*

Taha Mzoughi  
*Kennesaw State University, tmzoughi@kennesaw.edu*

*See next page for additional authors*

Follow this and additional works at: <http://digitalcommons.kennesaw.edu/facpubs>

 Part of the [Physics Commons](#), and the [Science and Mathematics Education Commons](#)

---

### Recommended Citation

Morris, G. A., Harshman, N., Branum-Martin, L., Mazur, E., Mzoughi, T., & Baker, S. D. (2012). An item response curves analysis of the Force Concept Inventory. *American Journal Of Physics*, 80(9), 825-831.

This Article is brought to you for free and open access by DigitalCommons@Kennesaw State University. It has been accepted for inclusion in Faculty Publications by an authorized administrator of DigitalCommons@Kennesaw State University. For more information, please contact [digitalcommons@kennesaw.edu](mailto:digitalcommons@kennesaw.edu).

---

**Authors**

Gary A. Morris, Nathan Harshman, Lee Branum-Martin, Eric Mazur, Taha Mzoughi, and Stephen D. Baker

## An item response curves analysis of the Force Concept Inventory

Gary A. Morris, Nathan Harshman, Lee Branum-Martin, Eric Mazur, Taha Mzoughi, and Stephen D. Baker

Citation: *American Journal of Physics* **80**, 825 (2012); doi: 10.1119/1.4731618

View online: <http://dx.doi.org/10.1119/1.4731618>

View Table of Contents: <http://scitation.aip.org/content/aapt/journal/ajp/80/9?ver=pdfcov>

Published by the [American Association of Physics Teachers](#)

---

### Articles you may be interested in

Erratum: "An item response curves analysis of the force concept inventory" [*Am. J. Phys.* **80**, 825–831 (2012)]  
*Am. J. Phys.* **81**, 144 (2013); 10.1119/1.4766939

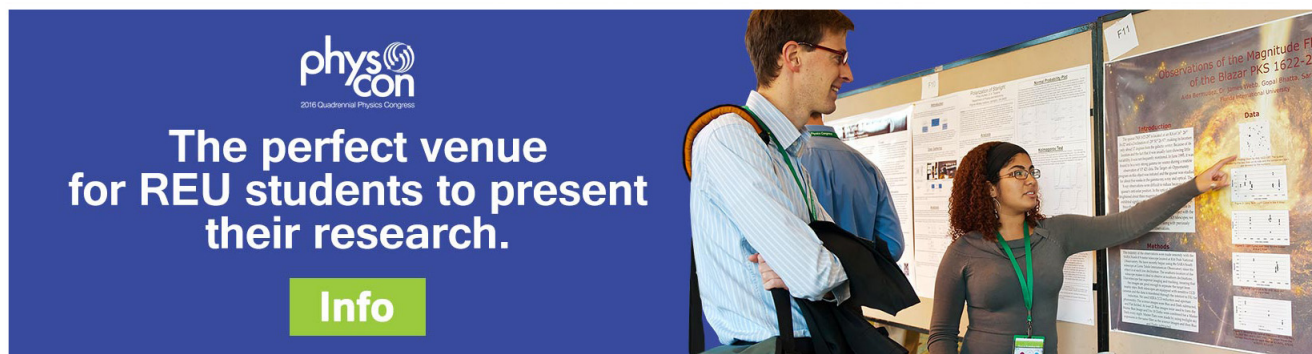
Comparing large lecture mechanics curricula using the Force Concept Inventory: A five thousand student study  
*Am. J. Phys.* **80**, 638 (2012); 10.1119/1.3703517

Analyzing force concept inventory with item response theory  
*Am. J. Phys.* **78**, 1064 (2010); 10.1119/1.3443565

Testing the test: Item response curves and test quality  
*Am. J. Phys.* **74**, 449 (2006); 10.1119/1.2174053

The effect of distracters on student performance on the force concept inventory  
*Am. J. Phys.* **72**, 116 (2004); 10.1119/1.1629091

---



**physcon**  
2016 Quadrennial Physics Congress

The perfect venue  
for REU students to present  
their research.

Info

# An item response curves analysis of the Force Concept Inventory

Gary A. Morris

*Department of Physics and Astronomy, Valparaiso University, Valparaiso, Indiana 46383*

Nathan Harshman

*Department of Physics, American University, Washington, DC 20016*

Lee Branum-Martin

*Texas Institute for Measurement, Evaluation, and Statistics, University of Houston, Houston, Texas 77204*

Eric Mazur

*Department of Physics, Harvard University, Cambridge, Massachusetts 02138*

Taha Mzoughi

*Department of Biology and Physics, Kennesaw State University, Kennesaw, Georgia 30144*

Stephen D. Baker

*Department of Physics and Astronomy, Rice University, Houston, Texas 77005*

(Received 30 November 2011; accepted 13 June 2012)

Several years ago, we introduced the idea of item response curves (IRC), a simplistic form of item response theory (IRT), to the physics education research community as a way to examine item performance on diagnostic instruments such as the Force Concept Inventory (FCI). We noted that a full-blown analysis using IRT would be a next logical step, which several authors have since taken. In this paper, we show that our simple approach not only yields similar conclusions in the analysis of the performance of items on the FCI to the more sophisticated and complex IRT analyses but also permits additional insights by characterizing both the correct and incorrect answer choices. Our IRC approach can be applied to a variety of multiple-choice assessments but, as applied to a carefully designed instrument such as the FCI, allows us to probe student understanding as a function of ability level through an examination of each answer choice. We imagine that physics teachers could use IRC analysis to identify prominent misconceptions and tailor their instruction to combat those misconceptions, fulfilling the FCI authors' original intentions for its use. Furthermore, the IRC analysis can assist test designers to improve their assessments by identifying nonfunctioning distractors that can be replaced with distractors attractive to students at various ability levels.

© 2012 American Association of Physics Teachers.

[<http://dx.doi.org/10.1119/1.4731618>]

## I. INTRODUCTION

One of the joys of physics is that, unlike many other fields of inquiry, right and wrong answers are often unambiguous. Because of this apparent objectivity, multiple-choice tests remain an essential tool for assessment of physics teaching and learning. In particular, the Force Concept Inventory (FCI) (Ref. 1) is widely used in physics education research (PER).

Tests are designed with specific models in mind. The FCI was designed so that the raw score measures (in some sense) the ability of "Newtonian thinking." This article compares two methods of test analysis based on different models of the functionality of the FCI. These methods have the unfortunately similar names of item response curves (IRC) (Ref. 2) and item response theory (IRT) (e.g., Ref. 3). By comparing these methods, we will show that the model behind the IRC analysis is more consistent with that envisioned by the FCI designers. Additionally, it is easier to use and its results are easier to interpret in a meaningful way.

Any method of test analysis requires the construction of a model to quantify the properties of the complex, interacting "population + test" system. For example, in order to detect correlations in the data, models often assume that each test-taker has a true but unknown "ability." In assessing students, educators expect ability to be strongly correlated with raw

total score on the test. Weighing each question equally in determining ability is a common model, but other methods, such as factor analysis or IRT, use correlations in the test response data to weigh the questions differently. Many standardized tests subtract a fraction of a point from the raw score to "penalize" guessing. This model assumes that the probability of guessing the right answer to a question does not depend on the specific question, that guessing any particular answer choice on a particular question is equally likely, and that guessing works the same way at all ability levels. Other, more sophisticated approaches, such as those in factor analysis or item response theory,<sup>3</sup> statistically define the probability of response in order to estimate potentially different weights for items. These approaches can be viewed as drawing information from the correlations among items to estimate appropriate weights. By the very nature of a multiple-choice test, all models presume that there exists an unambiguously correct answer choice for every item. But, a key difference in models is associated with how they handle the distractors.

IRC uses total score as a proxy for ability level, which implies that all items are equally weighed. In order to implement IRC analysis, item-level data are necessary. IRC analysis describes each item with trace lines for the proportion of examinees who select each answer choice, including the distractors, at each ability level (see Fig. 1). To make an IRC

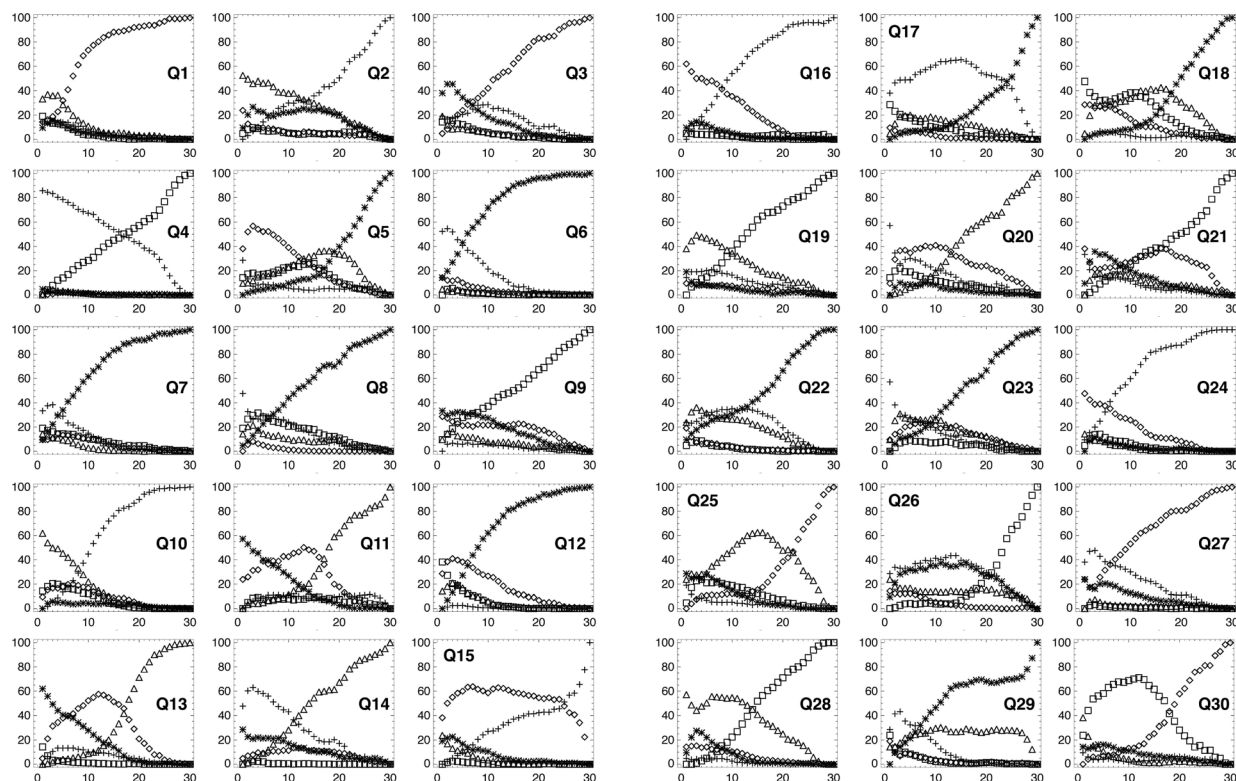


Fig. 1. The IRCs for all 30 questions on the FCI using the approach of Morris *et al.* (Ref. 2). This figure is analogous to Fig. 4 in Wang and Bao (Ref. 4). Each graph shows the percentage of respondents (vertical axis) at each ability level (horizontal axis) on a given item selecting a given answer choice: + = 1; \* = 2;  $\diamond$  = 3;  $\triangle$  = 4;  $\square$  = 5.

graph for a single item, we must separate the respondents by ability level and then determine the fraction at each ability level selecting each answer choice. For example, in our data set, there were 116 students of ability level 20 who answered Question 8. Of those, 81 selected Answer Choice 2, the correct answer. Thus, the trace line for the IRC of the correct answer (asterisks in Fig. 1) passes through the point (20, 69%). When repeated for all ability levels, all questions, and all answer choices, we arrive at the final product of IRC analysis: a set of traces for each item, such as those for the FCI appearing in Fig. 1.

IRT analysis derives student ability by a statistical analysis of item responses and does not necessarily weigh all questions equally. In fact, IRT uses a probabilistic model that is a convolution of examinee response patterns with a specific model of item functionality. The IRT analysis of Wang and Bao<sup>4</sup> employs the three parameter logistic (3PL) model to analyze the functionality of only the correct answer choice. The final product of the 3PL IRT model is an estimate for each item of three conceptually useful numbers: difficulty, discrimination, and guessing probability. While 3PL is one example of an IRT model, a variety of models with different algorithms and underlying assumptions have appeared in the literature.<sup>5,6</sup> One of the strengths of IRT analysis is that these parameters should be robust with respect to estimation from different populations.

We argue that, in the case of the FCI, IRC analysis yields information overlooked or not easily obtained with other methods and better reflects the intentions of the designers of the FCI. The distractors in the FCI are tailored to specific physical misconceptions that had been identified (by student interviews and instructor experiences) to exist in the population. That is, different distractors in an FCI item were

designed to function differently. The 3PL IRT analysis assumes that all wrong answers for a given item are equivalent. We will provide IRC analysis examples that show, in an easy-to-interpret graphical form, how different distractors appeal to students with varying abilities. If distractors are not all equally “wrong,” we could estimate ability in a more sensitive manner, effectively giving partial credit for some wrong answers—those that suggest a deeper understanding—than other wrong answers (cf. Bock<sup>7</sup>). Furthermore, we may assess the effectiveness of items and alternative answer choices as to how well they function in measuring student understanding and misconceptions.<sup>2</sup>

The IRC analysis is particularly meaningful in the case of the FCI because IRT analyses demonstrate the strong correlation between ability and total score for this test.<sup>3</sup> This quality justifies the use of raw score as a proxy for ability in the model that underlies IRC analysis.

Below we illustrate the power of IRC analysis in probing the performance of items on the FCI, with Sec. II comparing results from Ref. 2 with subsequent publications. Section III focuses on the differences between the 3PL IRT analysis and IRC analysis for the FCI and argues that IRC analysis is the more appropriate model for formative assessments like concept tests in PER. Section IV provides an in-depth analysis using the IRC approach for several FCI questions. We conclude in Sec. V.

## II. INSIGHTS FROM RECENT ANALYSES COMPARED WITH IRC ANALYSIS IN REF. 2

In Ref. 2, we outlined our IRC analysis approach to evaluating the effectiveness of test questions and applied that technique to three sample questions from the FCI. Our technique

provides a practical approach by which a broad cross-section of physics teachers could engage in a more thorough understanding of the functionality of items on the FCI specifically and multiple-choice tests more generally. Furthermore, our approach allows test designers to evaluate not only the efficacy of test items but also, within those items, the functionality of specific answer choices, both correct answer choices and distractors. Finally, we suggested the possibility for evaluating student knowledge level based not only upon the traditional dichotomous right/wrong scoring but also upon the specific incorrect answer choices selected.

Since the publication of Ref. 2, several studies have used IRT models to assess student performance and test item functionality, though to date, ours remains the only paper to examine the performance of the distractors within each item. Ding and Beichner<sup>8</sup> provide an overview of analysis techniques used to probe measurement characteristics of multiple-choice questions, including classical test theory, factor analysis, cluster analysis, and IRT. Schurmeier *et al.*<sup>9</sup> recently described a 3PL IRT analysis of General Chemistry assessments at the University of Georgia and found implications for the impact of item wording on question difficulty and discrimination. Marshall *et al.*<sup>10</sup> applied full-information factor analysis (FIFA) and a 3PL IRT model to 2003–2004 data from the Texas Assessment of Knowledge and Skills (TAKS) for science given to 10th and 11th grade students. This analysis indicated that the test was more a measure of generic testing ability and basic quantitative skills than specific proficiency in biology, chemistry, or physics and suggested the utility of a 3PL analysis in the design process for future assessments. (Marshall *et al.* also noted that Ref. 2 was similar to “item-observed score regression” described in Lord.<sup>11</sup> Indeed, we were unaware of this reference at the time of our original publication.)

Most relevant to our original work are the papers by Planinic *et al.*<sup>12</sup> and Wang and Bao.<sup>4</sup> Planinic *et al.* used a Rasch model (in effect, a one-parameter IRT model, distinguishing items by difficulty) to evaluate the functionality of the FCI for 1676 Croatian high school students and 141 Croatian college students. They identified problems with the functionality of some items on the FCI and made suggestions for improvement. Wang and Bao employed 3PL IRT analysis to the FCI. As was the case with the study by Marshall *et al.*, Wang and Bao used a 3PL model. The strength of 3PL IRT is the straightforward way in which it estimates difficulty and discrimination parameters by fitting the response data to a curve of probability of correct response plotted against ability. One limitation of such an approach is that, although 3PL IRT can model respondent guessing, it assumes that all distractors within an item function equally. We have demonstrated this not to be the case for the FCI. One way to overcome this weakness is to use a full-blown nominal IRT model in which response probabilities for each distractor are estimated, conditional on student ability. For example, the nominal response model<sup>7</sup> of Bock can be fit in MULTILOG software, but Bock’s parameters are not directly interpretable the way that the parameters “a,” “b,” and “c” of the 3PL in Wang and Bao are. Alternatively, the IRC approach can yield reasonably sophisticated analyses and is quite a bit easier to implement. Other approaches include the nonparametric approach of Ramsay’s TestGraf, a graphical approach to item analysis that makes fewer assumptions than standard IRT.<sup>10</sup>

For comparison with the approach of Wang and Bao and as an example to teachers and researchers looking for a prac-

Table I. The percentage of students who responded correctly to each question on the FCI.

Question	% Correct	Question	% Correct	Question	% Correct
1	71.6	11	25.7	21	32.9
2	34.6	12	65.2	22	42.1
3	51.5	13	26.3	23	39.6
4	37.1	14	39.5	24	66.6
5	19.2	15	29.0	25	21.4
6	73.6	16	59.2	26	13.2
7	66.4	17	17.6	27	59.4
8	50.4	18	22.2	28	36.7
9	45.9	19	45.6	29	50.8
10	54.0	20	32.3	30	25.8

tical approach to investigate item functionality, we applied IRC analysis to all 30 questions on the FCI. We used total score as a proxy for ability level, an approach supported by the findings of Wang and Bao—see their Fig. 5 and analysis that suggests a strong correlation between their IRT-based estimate of student ability ( $\theta$ ) and total test score:  $r^2 = 0.994$ . The high correlation between the total score and model-estimated ability does not mean that individual items are equivalent or that wrong responses to specific items are equivalent. We note that there may be value in the particular wrong answers given: experienced teachers know that some wrong answers are better (i.e., indicate a higher, though still imperfect state of understanding) than others. Using that information could provide the next logical step in diagnosing dysfunctional items or distractors and may lead to more reliable assessments of student ability levels (a benefit noted by Bock<sup>7</sup> in the nominal response model).

Figure 1 shows the IRCs for all 30 FCI questions from a database of >4500 student responses compiled by Harvard University, Mississippi State University, and Rice University, while Table I relates the percentage of students who answered each question correctly. This figure provides an analog to Fig. 4 in Wang and Bao. Some of the differences between the sets of graphs in the figures may be attributed to the fact that the data sets are independent of one another. The IRC analysis graphs emphasize the information in the distractors, which provides an instructor with an important diagnostic tool for evaluating student misunderstandings and item functionality, empowering instructors both to correct the most troublesome student misconceptions and develop more efficient test items.

In Ref. 2, we examined three questions from the FCI in some detail. First, we identified Question 11 as a difficult (25.7% correct) but an efficient question, with an IRC for the correct answer choice showing a sharp slope near an ability level/total score of 17 (out of a possible 30), allowing for a clear discrimination between students above and below this ability level, a result attainable with the 3PL IRT analysis as well. However, using IRC, we also identified that Answer Choice 2 was particularly popular among lower ability students, while Answer Choice 3 was most popular among students in the middle of the ability range. Interestingly, Planinic *et al.* identify this question as problematic in that their single-parameter Rasch model fails to predict the students who will answer this question correctly.<sup>12</sup> While their sample is different from ours, their model assumes that the low-ability students who get this question correct have done

so by guessing, yet our IRC analysis suggests a very low percentage of low-ability students are getting the correct answer; they are much more likely to select one of two alternative, yet attractive, incorrect answer choices. Like Wang and Bao, the analysis by Planinic *et al.* does not examine or make use of the set of incorrect answer choices.

Next, we considered Question 9, which Wang and Bao rated as the sixth hardest question on the FCI but the second most likely question on which to guess correctly.<sup>4</sup> IRC analysis reveals that this unlikely pairing occurs (at least in our sample) because Answer Choices 1 and 4 do not function very well, attracting few students of any ability level (the curves are low and have shallow slopes). Our analysis is consistent with that of Planinic *et al.*, who placed this question in the middle of their difficulty scale.<sup>12</sup>

Finally, we identified Question 4 as moderately difficult (37.1% correct), but very inefficient, since three of the answer choices were not functioning at all, each attracting less than 2% of the students. Wang and Bao ranked this as the ninth hardest question and tenth easiest to guess correctly.<sup>4</sup> Planinic *et al.* rated the difficulty of this question in the middle of the range.<sup>12</sup> As we see in these examples, the additional information provided by IRC analysis (e.g., the functionality of individual answer choices, the attractiveness of individual answer choices to different subsets of the population segregated by ability level, etc.) could lead to greater insights into the actual ability levels of students taking the test and the functionality of test items.

### III. A COMPARISON OF IRC ANALYSIS WITH 3PL IRT ANALYSIS FOR THE FCI

Let us now explore further a comparison between the 3PL IRT model analysis by Wang and Bao and our approach for the FCI. The 3PL IRT model of Wang and Bao naturally produces estimates of three different parameters, labeled  $a$ ,  $b$ , and  $c$ , which correspond to discrimination, difficulty, and guessing, respectively.

First, Wang and Bao estimate the discrimination of an item on the FCI from the 3PL model parameter  $a$ , which is essentially the slope of the 3PL model fit for the correct answer choice at the ability level corresponding to 50% correct response. In the IRC analysis, we can qualitatively judge the discrimination of an answer choice from the slope of our curves: the steeper the slope, the more discriminating the answer choice. Table I in Wang and Bao indicates that Questions 5, 13, and 18 are the most discriminating items on the FCI. Examining our Fig. 1, we see that the IRCs for the correct answer choices in each of these cases are each characterized by a steep slope.

The IRC graphs in Fig. 1, however, illustrate the limitation of the estimation of item discrimination from the 3PL IRT analysis. In particular, the 3PL IRT model assumes that all distractors function equally. This assumption was explicitly not made in the design of the FCI, and Fig. 1 demonstrates that this assumption is not supported by the data. The behavior of the IRCs for the FCI answer choices indicates that a more sophisticated model than the 3PL IRT is needed. For example, Question 29 is rated by Wang and Bao as the least discriminating item on the FCI.<sup>4</sup> Yet, when we examine the IRC for the correct answer choice, we find three regions of ability level with different levels of attraction to this answer choice: Below a total score of 12, the probability of a

student's selecting the correct answer choice increases linearly with a moderate slope; for a total score of 12 to 27, the probability is nearly constant, increasing only slightly with ability level; and from 28 to 30, the probability rises rather sharply. How to properly characterize the discrimination of such an item remains an open question, but the behavior of the IRC suggests that the 3PL model may overlook some information in that the performance of examinees may not be a smoothly increasing function of ability level.

Furthermore, the content of Question 29 extends beyond the Newtonian thinking dimension that the FCI is designed to measure in that the influences of the forces of gravity, air resistance, and buoyancy are included in the answer choices, while instruction at the time of the administration of the FCI may not have addressed the last two topics, even at the time of the posttest.

In general, however, given the shape of the fitting functions used in the 3PL model, only IRCs that monotonically increase and have at most a single inflection point can be modeled. From Fig. 1, we see that the 3PL IRT model likely will have problems fitting the data not only from Question 29 but also from Questions 11 and 15. Questions 2, 5, 8, 14, 17, and 20 also may cause difficulties for 3PL IRT based on our criteria. With our sample size, however, a local variation in the shape of the IRC to a specific answer choice of less than ~5% may not be statistically significant. We need a larger data set to investigate these questions more carefully. In any case, 3PL IRT results for these items should be examined carefully in light of these observations.

Second, Wang and Bao describe item difficulty using parameter  $b$ . Figure 2 shows a scatter plot of percent correct versus the difficulty parameter of Wang and Bao—the linear fit has a correlation coefficient of 0.89. The good agreement suggests that using percent correct on an item on the FCI provides quite a good estimate of item difficulty. Wang and Bao specifically identify Questions 1 and 6 as easy.<sup>4</sup> We concur, finding 71.6% and 73.6% of student examinees, respectively, responded with the correct answer choices—these being the two highest percentages of correct responses. The IRCs also allow some insight as to why these questions are at the easy end of the spectrum on the FCI. In both cases, one answer choice is attractive to students at the low end of the total score axis. In the case of Question 1, it appears that students with low total scores show a slight preference for Answer Choice 4 (which corresponds to the misconception that heavier objects fall faster). In the case of Question 6, two of the distractors are not functioning well at any ability level (Answer Choices 4 and 5 appear as simply more extreme and unlikely versions of Answer Choice 3 as related to the subsequent path of an object freed from circular motion). We discuss both of these examples in more detail in Sec. IV below.

Wang and Bao identify Questions 25 and 26 as the most difficult, although by the data in their Table I, the four most difficult questions (in order) are Questions 17, 26, 15, and 25. Using our analysis, we found rates of correct responses of 17.6%, 13.2%, 29.0%, and 21.4%, respectively. We also found Question 5 to be very difficult with only 19.2% correct; Wang and Bao found this to be the sixth most difficult question in their analysis. In the case of Question 15, the difficulty mainly results from one well-functioning and attractive distractor: Answer Choice 3 (if the car pushes the truck forward, it must be exerting a greater force on the truck than vice versa). The difficulty associated with Question 17, in

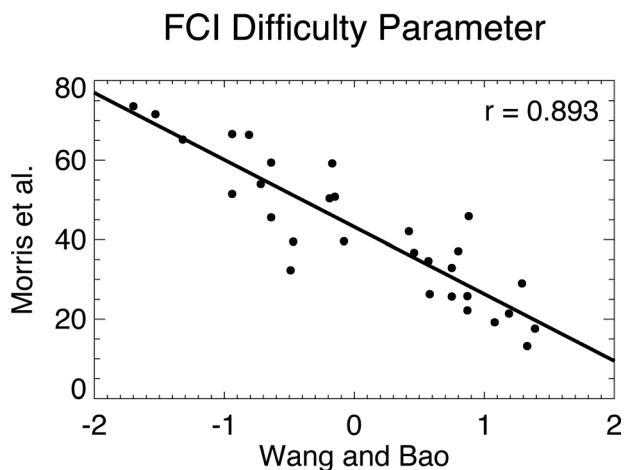


Fig. 2. A comparison of the percent correct (this work, vertical axis) with the difficulty parameter (“b,” horizontal axis) from Wang and Bao (Ref. 4).

fact, comes from the very same misconception. Here, Answer Choice 1 is attractive across ability levels and is associated with the misconception that, in order to move up, even at constant speed, the force up must be greater. Thus, the IRC analysis allows us to better characterize why questions are relatively easy or difficult.

Third, Wang and Bao estimate the guessing probability or parameter  $c$ . Estimates of the guessing parameter are notoriously noisy and difficult to make because the low end of the ability scale is usually sparsely populated. If we examine the 3PL curves in Fig. 4 of Wang and Bao, we find that, when their model and data disagree, the model systematically overestimates the percent correct as compared to the data at the low end of the scale (conversely, although less relevant here, we note that at the high end of the scale, the model typically underestimates the percent correct at each ability level). Part of the difficulty once again may be that the 3PL IRT model assumes that all the incorrect answer choices are functioning equally, an assumption that we are not required to make with IRC analysis and that is not valid in the case of the FCI. As an alternative, one might estimate the fraction guessing from the IRC analysis by simply extrapolating the correct IRC to a “0” ability level. We note, however, that if someone with no knowledge in the content areas covered by the FCI took the test, we would expect the guessing fraction to be 20%, given that five answer choices appear with each question. The fact that, in our Fig. 1, we never find an IRC with greater than 12% correct at 0 ability level communicates clearly that the distractors have been chosen carefully to attract students with some knowledge of specific and common misconceptions. That said, all of these approaches suffer from the limited data at the extremely low end of the scale, so we have little confidence in any estimates of a guessing parameter.

We provide a final point to further highlight the additional information that is provided from our IRC analysis compared to a dichotomous IRT analysis like 3PL, which examines only the correct answer choices. Figure 3 shows a “dichotomously” constructed IRC diagram for Question 13, with curves only for “correct” and “incorrect,” the latter consisting of the sum of the curves from all the incorrect IRCs as well as the no response data. This figure is the analog to all of the graphs in Fig. 4 from Wang and Bao. A comparison of our Fig. 3 with the corresponding graph in our Fig. 1 reveals a substantial loss

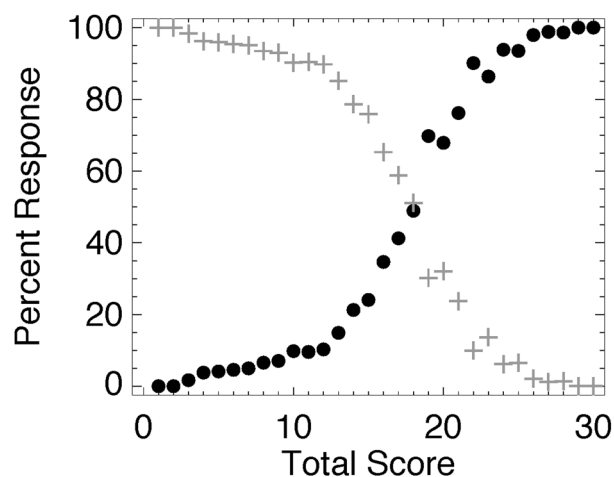


Fig. 3. IRCs for a “dichotomously” scored Question 13 from the FCI.

of information content. This question has two well-functioning distractors that attract students at distinct ranges of ability. The effectiveness of these distractors is lost in the analysis of Fig. 3, in which we can only see the discrimination of the correct answer choice between students scoring above and below a total score around 18.

#### IV. ADDITIONAL FCI INSIGHTS FROM IRC ANALYSIS

As demonstrated above, the IRC analysis allows for a rich investigation of the functionality of items and particular answer choices within items on assessment instruments. Reference 2 provided a characterization of three items on the FCI (reviewed in Sec. II above). Here, we demonstrate the capabilities of the IRC approach more completely by examining several additional items in further detail. The analysis below is not meant to be definitive or conclusive; a still larger data set and separation of pre- and postinstruction results would be helpful in this regard. However, it is presented to give the reader better insights into the kinds of analysis that the IRC approach permits, especially if combined with student interviews.

*Question 1:* This question (dropping two metal balls of different masses from a roof top) allows the examiner to discriminate at the lower end of the total score range, with the IRC for the correct answer choice crossing 50% correct rather steeply near a total score of 7. A student with even minimal previous physics instruction will have had direct engagement with this point and will select correct Answer Choice 3 (both hit the ground at the same time). Perhaps because it appeals to naïve, Aristotelian intuition about motion, Answer Choice 4 (heavier ball hits the ground first) seems to attract students at the lowest end of the total score range, while the other three answer choices do not seem to attract students at a rate higher than guessing. Improvement of this question could be made by changing one of the low-appeal distractor choices (1, 2, or 5) so that it would be more attractive to moderate ability students. Finding such an alternative answer choice might be accomplished by using FCI scores to identify moderate ability students, interviewing them to identify more subtle gradations of misconceptions about weight and acceleration, and then reworking the entire question. Further IRC analysis could allow an instructor to



identify whether such an answer choice had been effectively constructed.

*Question 5:* The IRCs for this item (forces acting on a ball traveling around a horizontally oriented semicircular track) indicate that all of the answer choices are functioning reasonably well. The IRC for the correct answer choice crosses 50% correct near a total score of 22, suggesting that the correct answer discriminates between students in the high and moderate ability ranges. As for the distractors, Answer Choice 1 (only gravity acts) is attractive to students at the very lowest end of the total score range ( $<2$ ). Answer Choice 3 (gravity and a force in the direction of motion act) preferentially attracts students at the low end of the range as well (2–10). The IRC for Answer Choice 5 (gravity, a force in the direction of motion, and a force pointing from the center of the circle outward) shows a broad peak in the range of total scores of 9–16, while the IRC for Answer Choice 4 (gravity, a force exerted by the channel toward the center of the circle, and a force in the direction of motion) shows a broad peak in the total score range of 14–23. This item provides the only example in this data set for which all five answer choices seem to be functioning at some level, perhaps because students must put together several concepts from Newton's laws and circular motion in order to select the correct answer choice.

*Question 6:* Examining the content of this item, we see this question probes student understanding of Newton's first law by asking what happens to the same ball traveling around a horizontally oriented semicircular track as in Question 5, but as the ball exits one end. This question ranked as one of the easiest on the FCI by all analyses (Wang and Bao, Planinic *et al.*, and ours). Nevertheless, this item does provide some discrimination at the bottom end of the ability scale, with Answer Choice 1 being the most attractive to those of low ability. Answer Choice 1 corresponds to the ball continuing to move in a circular path. Students who select this answer choice do not understand that circular motion requires the presence of a net force. Answer Choices 3, 4, and 5 are all variants of the same misconception—an expression of the “centrifugal force” causing objects moving along a circular path to feel a false force away from the center. The IRCs indicate that these distractors are not functioning particularly well. Since the slight differences in the shapes of the curves for these three answer choices have no motivation, a skilled test taker would understand that none of these choices can be correct. This item, therefore, might better be posed with three answer choices rather than 5, eliminating two of Answer Choices 3 through 5.

*Question 13:* This question examines the forces present on a ball thrown directly upward after it has been released but before hitting the ground. The IRCs for three answer choices on this item suggest that they could be used to characterize the ability level of a student into one of three ranges. Answer Choice 2 (varying magnitude force in the direction of motion) is preferred by a plurality of students with total scores  $<5$  and by more than about half of respondents with total scores  $<3$ . Students selecting this answer choice, therefore, have a high probability of being at the low end of the ability range. Answer Choice 3 (gravity plus a decreasing force in the direction of motion on the way up only) is preferred by a plurality of students with  $7 < \text{total score} < 17$  and by more than about half of respondents with  $10 < \text{total score} < 16$ . Students selecting this answer choice, therefore, are likely to be in the middle of the ability range. Finally,

Answer Choice 4, the correct answer, is preferred by students with total scores  $>17$ . The sharp slope of this IRC identifies this item as highly discriminating. Thus, by examining the response to this one item, it is possible to roughly gauge the ability level of a student respondent. In conjunction with an analysis of other responses (correct and incorrect), the examiner increases his/her probability of correctly identifying the respondent's true ability level. We note that, in general, this ability level may or may not match the student's total score; it is possible for a student with a lower total score to be identified as having a higher ability level based on the specific set of incorrect and correct answer choices selected. Thus, we can develop a more accurate assessment of student ability level using an IRC analysis. Furthermore, the results related to this particular item on the FCI suggest different instructional strategies may be needed to address differing misconceptions students have regarding the subject matter. In particular, students at the low end of the ability scale need to learn why Answer Choice 2 is incorrect (perhaps the students are confusing the force of gravity with velocity), while students of middle ability need to learn why Answer Choice 3 is incorrect (perhaps these students understand the force of gravity but also describe a fictional upward force that closely correlates to momentum). We would need to conduct student interviews to probe these misconceptions more carefully. Nevertheless, in this case, IRC analysis has revealed a possible strategy for tailored instruction (e.g., online or computer assisted learning activities).

*Question 15:* This question examines the forces as a car pushes a truck while accelerating. The IRCs for this item indicate that, as written, students are primarily attracted to incorrect Answer Choice 3 (force of car on truck is greater than force of truck on car) across all ranges of ability. Experienced teachers are not surprised by this result, even though Answer Choice 1 is a nearly direct translation of Newton's third law into this scenario. The IRC for the correct Answer Choice 1 crosses 50% correct at a very high total score (ability) of 25, making this a very difficult item. For the instructor, it would seem that the primary duty in this case is to correct the misconception students hold in responding with Answer Choice 3: Newton's third law is not a “sometimes” law, it is an “all-the-time” law (except when there is momentum transfer to a field, but that exception only highlights its usefulness). Of course, when the instructor succeeds in correcting this misconception, the shapes of the other IRCs may well change, forcing the instructor to target future class time in response to those changes. As with the discussion of Question 13, we find the prospects for customized instruction aided by IRC analysis of assessment instruments to be exciting.

## V. CONCLUSIONS

This paper presents an overview of the types of analysis that are possible using IRCs and was inspired by the recent paper by Wang and Bao, which uses IRT to evaluate the FCI.<sup>4</sup> In principle, the parameters extracted from IRT can be useful for scoring tests on a more sensitive metric, but in the case of the FCI, there is such strong correlation between ability and total score that this is not required for meaningful analysis.

The purpose of the 3PL IRT model described in Wang and Bao is to illustrate IRT and provide parameters to characterize the FCI test items. By contrast, the purpose of the IRC approach is to illustrate item and distractor diagnostics in a

way usable by classroom teachers and applicable to diagnostic instruments beyond the FCI. In a qualitative comparison of IRC analyses of the FCI with the 3PL IRT model results, we find that we are able to characterize item difficulty by using percent correct, which correlates strongly with Wang and Bao's difficulty parameter,  $b$ . However, we claim that the discrimination parameter,  $a$ , and the guessing parameter,  $c$ , as extracted by 3PL IRT model in Wang and Bao do not capture the variable functionality of distractors with ability level. Furthermore, we demonstrate using the IRC analysis that there are several questions on the FCI for which the 3PL IRT model is not likely to produce a good fit to the data.

The strength of the IRT model is that the parameters that are derived are more reliably sample-independent. This means that item parameters estimated from a high-performing representative sample of students would match those estimated from a low-performing representative sample of students. Another strength of IRT is that person and item estimates are model-based and can therefore be evaluated for fit or validity.

The strength of the IRC analysis, on the other hand, is that it provides a simple, descriptive, graphical approach to evaluating the performance of distractors and for defining student ability levels in relation not only to dichotomous scoring but also to the set of incorrect answers that they select. For the current paper, we suggest that IRC can be informative to classroom instructors. IRC does not require extensive psychometric training or specialized software. However, we look forward to future research in which the nominal response model<sup>7</sup> can be applied to take advantage of the benefits of IRT while examining the potentially different

information that was designed into the distractors on the FCI. While alternative models might not make a great difference in scoring the FCI, close examination of item performance will help instructors and researchers in physics.

- <sup>1</sup>D. Hestenes, M. Wells, and G. Swackhamer, "Force concept inventory," *Phys. Teach.* **30**, 141–158 (1992).
- <sup>2</sup>G. A. Morris, L. Brnum-Martin, N. Harshman, S. D. Baker, E. Mazur, S. Dutta, T. Mzoughi, and V. McCauley, "Testing the test: item response curves and test quality," *Am. J. Phys.* **74**, 449–453 (2006).
- <sup>3</sup>R. P. McDonald, *Test Theory: A Unified Treatment* (Erlbaum, Mahwah, NJ, 1999).
- <sup>4</sup>J. Wang and L. Bao, "Analyzing force concept inventory with item response theory," *Am. J. Phys.* **78**, 1064–1070 (2010).
- <sup>5</sup>D. Thissen and L. Steinberg, "A taxonomy of item response models," *Psychometrika* **51**, 567–577 (1986).
- <sup>6</sup>S. E. Embretson and S.P. Reise, *Item Response Theory for Psychologists* (Erlbaum, Mahwah, NJ, 2000).
- <sup>7</sup>R. D. Bock, "Estimating item parameters and latent ability when responses are scored in two or more nominal categories," *Psychometrika* **37**, 29–51 (1972).
- <sup>8</sup>L. Ding and R. Beichner, "Approaches to data analysis of multiple-choice questions," *Phys. Rev. ST Phys. Educ. Res.* **5**, 020103 (2009).
- <sup>9</sup>K. D. Schurmeier, C. H. Atwood, C. G. Shepler, and G. J. Lautenschlager, "Using item response theory to assess changes in student performance based on changes in question wording," *J. Chem. Educ.* **87**, 1268–1272 (2010).
- <sup>10</sup>J. A. Marshall, E. A. Hagedorn, and J. O'Connor, "Anatomy of a physics test: validation of the physics items on the Texas Assessment of Knowledge and Skills," *Phys. Rev. ST Phys. Educ. Res.* **5**, 010104 (2009).
- <sup>11</sup>F. M. Lord, *Applications of Item Response Theory to Practical Testing Problems* (Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, 1980).
- <sup>12</sup>M. Planinic, L. Ivanjek, and A. Susac, "Rasch model based analysis of the Force Concept Inventory," *Phys. Rev. ST Phys. Educ. Res.* **6**, 010103 (2010).

### MAKE YOUR ONLINE MANUSCRIPTS COME ALIVE

If a picture is worth a thousand words, videos or animation may be worth a million. If you submit a manuscript that includes an experiment or computer simulation, why not make a video clip of the experiment or an animation of the simulation. These files can be placed on the Supplementary Material server with a direct link from your manuscript. In addition, video files can be directly linked to the online version of your article, giving readers instant access to your movies and adding significant value to your article.

See <http://ajp.dickinson.edu/Contributors/EPAPS.html> for more information.